Identify Potential Deposit Subscribers of The Portuguese Retail Bank Market

Contents

1 Introduction **2**

1.1 Background 2

1.2 Statistical Objective 2

2 Exploratory Analyisis **2**

3 Binary Logistic Regression Model **4**

3.1 General Model Form and Variable Selection 4

3.2 Model Assumptions 4

3.3 Model Validation 4

3.4 Model Diagnostics 5

4 Random Forest (RF) Model **6**

4.1 General Model Form and Variable selection 6

4.2 Model Assumptions 6

4.3 Model Validation 6

5 Model Comparison **7**

6 Conclusions, Limitations, and Future Research **7**

7 References **8**

8 Appendix **9**

Team ID: 4

Kenneth Broadhead (Coding + Editing)

Koral Buch (Coding + Editing + Submitting) (Group Leader)

Min Kim (Group Setting + Editing)

Nanhao Chen (Coding + Editing)

Github Repository (Click me!)

1 Introduction

1.1 Background

A Portuguese retail bank initiated a telemarketing campaign from 2008 to 2013 aiming to maximize the subscription of new clients to a long-term deposit. This campaign used a direct method of marketing through cellphone or telephone. A subset of the data, of the years 2008 to 2010, was uploaded in February 2012 to the UC Irvine Machine Learning Repository and publicly available for research purposes [1].

1.2 Statistical Objective

The report focuses on the construction of a model for predicting whether or not a retail banking telemarketing campaign is successful in Portugal. In order to investigate and inform any crucial information of the dataset to the clients who are interested in this market, summary statistics using visual representations are included and explained. To predict the success or failure of telemarketing for subscription to a long-term deposit, predictive models such as logistic regression and random forests are utilized [2]. Corresponding model diagnostics and comparisons of the models' performance are discussed.

2 Exploratory Analyisis

Since our ultimate goal is constructing predictive models, we utilize the full data set, with all 41,188 observations, avialable at the repository. This provides us with all potentially relevant predictors, as well as enough data to split into training and validation sets. Due to the large size of the data set, we split the data (randomly) 50/50 into training and validation sets [3]. Initial exploration of the data shows there are several categorical predictors that have 'unknown' as a level. Since this lack of knowledge could be potentially useful to bank telemarketers, we treat these missing values as factors.

We removed the "duration" variable, since the duration of a call between a bank and client isn't known in advance, and would be unhelpful in building a predictive model. Additionally, we removed the variables concerning personal loans ("loan") and employment variate rate ("emp.var.rate"), for they cause extreme collinearity among the predictors. Finally, we removed the variable concerning defaulting on credit ("default"), for it is extremely unbalanced (only 3 'yes' values), resulting in instability in our logistic regression and random forest models.

Below we provide a few summary plots of interest (Figure 1 and 2). The frequency bar charts show some interesting behavior. The job plot shows that students and retired individuals are more likely than others to subscribe to a long-term deposit. The plot of marital status shows that each class of individuals is roughly equally likely to subscribe to a long-term deposit, with single and unknown individuals appearing to have a very slightly increased chance of subscribing.

In the conditional probability plots, the relationships between the response and the predictors age, the consumer price index (CPI), and the consumer confidence index (CCI) are examined. A clear pattern is shown in the plot involving age: as the age of the client increases, so does the probability that the client subscribes to a long-term deposit. No clear patterns are seen in the CCI and CPI plots, but there are pockets of increased probability of subscription, possibly indicating more complex behavior that could be exploited for predictive purposes in conjunction with other predictors. More advanced predictive methodologies are outlined below.

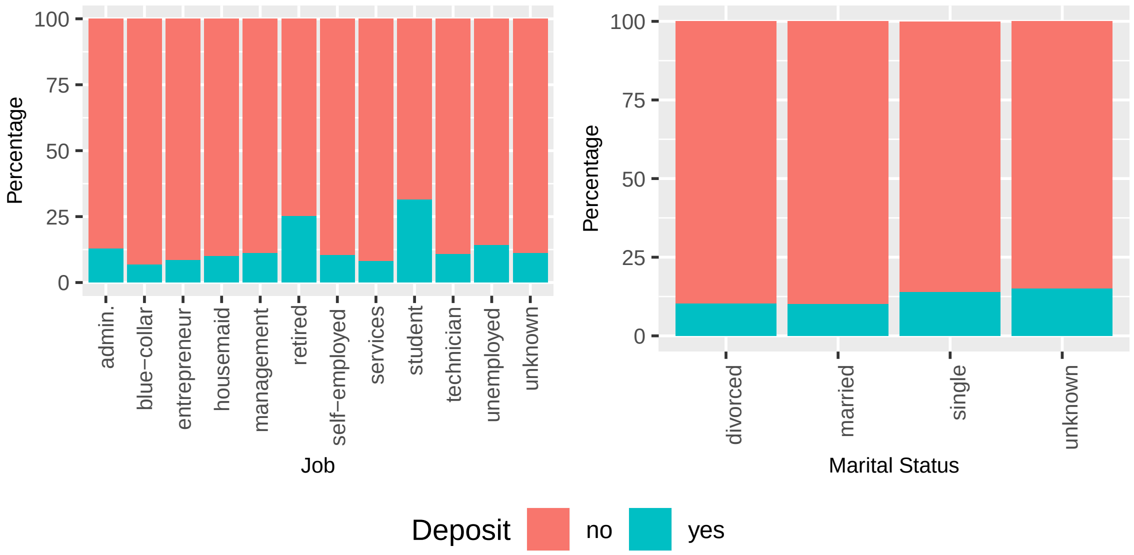

Figure 1: Categorical Stacked Bar Plots

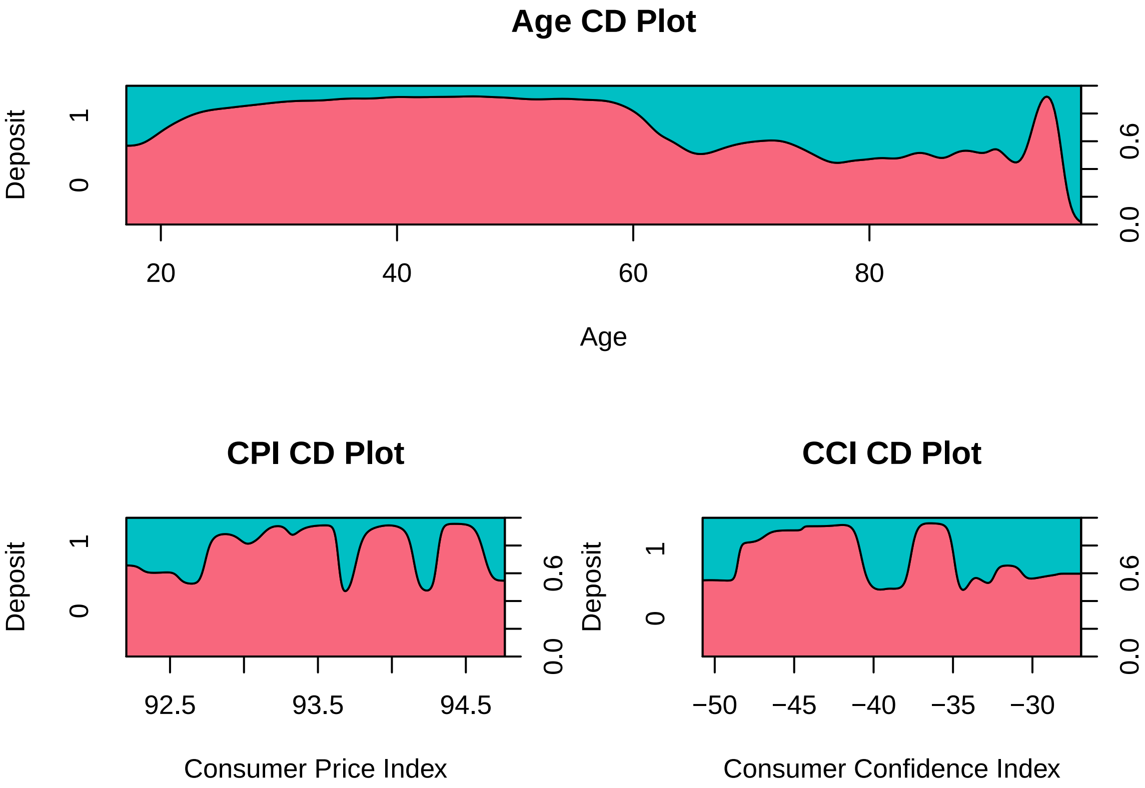

Figure 2: Conditional Density Plots

3 Binary Logistic Regression Model

3.1 General Model Form and Variable Selection

The binary logistic regression model is:

$ln(\frac{p}{1-p}) = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$

Where:

Y is the binary response variable, $Y = 1$ means the client subscribed for a deposit, and $Y = 0$ means the client did not subscribe for a deposit;

p is the probability that $Y = 1$;

b_0 is the interception at y-axis;

$x_1, ...x_n$ are the predictor variables;

$b_1, ...b_n$ are the regression coefficients of $x_1, ...x_n$, respectively.

In order to improve the prediction capabilities of the model, we fit several different models until we maximized the performance, as measured by the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, and the Matthews Correlation Coefficient (MCC). First, we fit a linear additive model with all 16 variables. Then, we added reasonable interaction and quadratic terms. While an exhaustive search for important second order effects was not feasible, we found that the addition of the following reasonable terms gave a model with the best prediction performance: Quadratic terms for age ("age"), the number of days that passed before a client was last contacted ("pdays"), the consumer confidence index ("cons.conf.idx"), and the Euro three-month Interbank Offered Rate ("euribor3m"). Additionally, an interaction term between age and marital status ("age" · "marital") and an interaction term between education and the consumer price index ("education" · "cons.price.idx") was found to improve model performance.

3.2 Model Assumptions

The assumptions for a logistic regression model are:

- Assumption of Appropriate Outcome Structure - For the binary logistic regression, the type of the dependent variable (outcome) should be binary. In case of the dataset we analyze, we build a binary logistic regression model since we are interested in a predictive model for binary response variables (Subscription: Yes, No).
- Assumption of Independent Observations - Logistic regression requires all observations to be independent of each other.
- Assumption of Absence of Multicollinearity - Logistic regression requires the independent variables to be not highly correlated with each other.
- Assumption of linearity of Independent variables and Log Odds - Logistic regression requires that the independent variables are linearly related to the log odds.

3.3 Model Validation

To validate our logistic regression model's predictive capabilities, we first fit the logistic regression model to the training data set. We then use this fitted model to make predictions based on the validation data set. Below we provide summaries of the model's performance in and out of sample performance. Table 1 and 2 show confusion matrices for the model's performance in the training data set (table 1) and the validation set (table 2). Note the strikingly similar performance. Furthermore, Figure 5 in the Appendix shows ROC curves for the model's performance in the training and validation data sets. Note the remarkable similarity of the two curves. The AUC for each curve is 0.7962 for the training data set, and 0.7907 for the validation set. The MCC for the training set is 0.372, while the MCC for the validation set is similar, at 0.332. The similar performances of the fitted logistic regression model in these two data sets suggests that these measures for performance accurately characterize the predictive performance of the logistic regression model. We thus fit a final logistic regression model using the full data set and proceed to model diagnostics.

Table 1: Confusion Matrix For Training Set

Target	Prediction	
	No	Yes
No	17983 (87.3%)	1734 (8.4%)
Yes	289 (1.4%)	588 (2.9%)

Table 2: Confusion Matrix For Validaiton Set

Target	Prediction	
	No	Yes
No	17944 (87.1%)	1788 (8.7%)
Yes	332 (1.6%)	530 (2.6%)

3.4 Model Diagnostics

We note that our first assumption (Appropriate Outcome Structure) is trivially satisfied, for we have binary response data. Furthermore, observations are independent of one another, as clients were contacted individually of one another. We addressed potential problems of collinearity in chapter 3.1 above. Finally, to investigate the linear relationship of independent variables and Log Odds assumption, we examine a plot of the residuals against a plot of the linear predictor. If the overall model is correct, a Lowess smooth of the plot should approximate a horizonal with zero intercept. Figure 3 below shows such a plot, with a Lowess smooth roughly approximating a horizontal line with zero intercept. Thus, the independent variables appear roughly linearly related to the log odds. This suggests the overall appropriateness of the fitted logistic regression model.

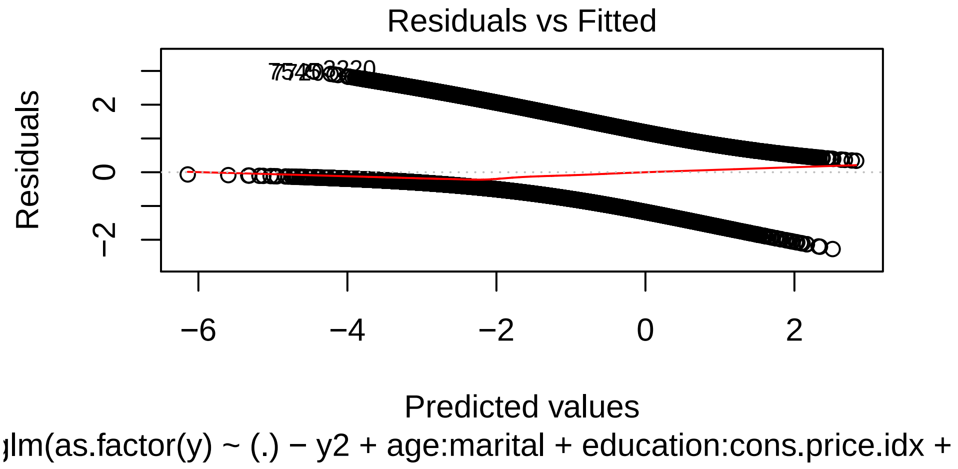

Figure 3: Residual Plot

4 Random Forest (RF) Model

4.1 General Model Form and Variable selection

To obtain a better prediction model, the random forest method was applied to train the model (RF Model). The random forests model is constructed by an ensemble of classification or regression decision trees. The model uses the random feature selection in the tree induction process and makes the prediction by cumulating the predictions of the branches. In general, the random forests model is fast to calculate, comparing to the other complex machine learning algorithms, and is as good as the best supervised learning algorithms. At the same time, the random feature selection in the random forests model makes this model less possible to overfit the data. Although the depth of the random forests method results in the difficulty of the data interpretation, this method can give us a good model with relatively low cost. Herein, the random forests model is constructed based on the same variable options as the logistic regression model used above in order to compare their performance.

4.2 Model Assumptions

The random forests method usually requires the balanced dataset, because the unbalanced dataset makes this method bias to the same direction. Since the response variable in our dataset is completely unbalanced, the class weight is used to re-balance this variable. Herein, the 'NO' and 'YES' classes are weighted inversely proportional to how frequently they appear in the dataset.

4.3 Model Validation

Table 3: Confusion Matrix For Training Set

Target	Prediction	
	No	Yes
No	17937 (87.1%)	0 (0%)
Yes	335 (1.6%)	2322 (11.3%)

Table 4: Confusion Matrix For Validaiton Set

Target	Prediction	
	No	Yes
No	17944 (87.1%)	1788 (8.7%)
Yes	332 (1.6%)	530 (2.6%)

Similarly, the confusion matrix of the RF Model was built in Table 3 and 4. The RF Model has a good performance on the training dataset with the AUC value as high as 0.998. Given to the prediction ability, the AUC value of the validation set is about 0.74, which is a little smaller than that of the training set (ROC curve in Figure 6 in the Appendix). Besides, the MCC values for the training set and validation set are 0.92 and 0.31 respectively. This difference between the training set in the RF Model may be caused by the unbalance of the dataset even though the re-weighted parameters have been considered when setup the model.

According to the importance variable analysis, the way to contact the clients ('contact' variable) plays an important role (11%) in the RF Model. Based on the whole dataset, the people contacted by cellular have more chance to subscribe a term deposit than those contacted by telephone. Besides, the 'euribor3m' (11%), describing the Euribor 3-month rate, and the 'age' of the clients (13%) are another two important variables in the dataset. It is reasonable that the investment behavior is associated with the loan interest rate as well as the age of the clients. The low loan interest rate (or even negative) will encourage the clients to do the investment, and at the time, older people probably tend to have more money to purchase the financial product. There are other important variables, such as the number of contacts to the client ('campaign', 8%), and the 'education' (10%).

5 Model Comparison

As noted above, the logistic regression model performed similarly in and out of sample. Thus, the final logistic regression model (fit using all data) measures of performance (MCC, ROC, AUC) likely provide an accurate measure of its predictive performance. However, the random forest model showed some notable discrepancies in performance. The measures of predictive accuracy for obtained using the validation set are therefore more accurate measures of its overall predictive performance. We therefore compare the performance of the final logistic model to the performance of the random forest on its validation data set.

The final logistic regression model had a MCC of 0.347, while the random forest had a MCC of roughly 0.31. Additionally, the final logistic model had an AUC value of 0.795, while the random forest had an AUC value of roughly 0.74. This difference is further highlighted by the joint ROC curves plotted blow in Figure 4. This discrepancy between the to models appear to be caused by overfitting on part of the random forest model. The training and validation sets support this conclusion, as one would expect predictive performance to be vastly superior in the training data versus the validation data when a model suffers from overfitting.

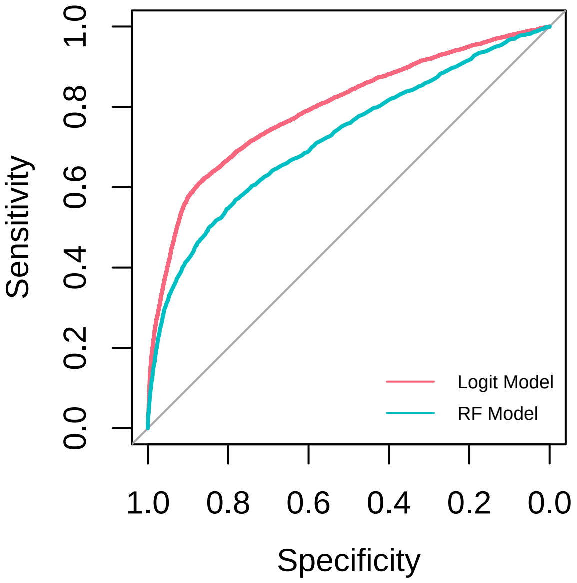

Figure 4: ROC Plot Logistic Regression Model and Random Forest Model

6 Conclusions, Limitations, and Future Research

In conclusion, while the random forest model performed very well in sample, the logistic regression model had slightly better out of sample performance across all measures. Thus, we think the logistic model a superior model for predictive performance, and recommend its use over the random forest model.

7 References

1. Center for Machine Learning and Intelligent Systems: UCI Machine Learning Repository, Bank Marketing Data Set. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

2. S. Moro, P. Cortez and P. Rita., 2014. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31.

3. Kutner, M.H., 2005. Applied Linear Statistical Models, 5th.

8 Appendix

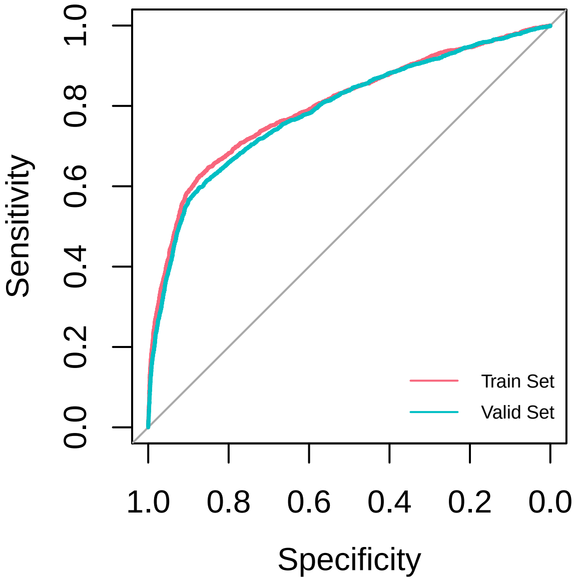

Figure 5: ROC Plot For Training and Validation Datasets of Logistic Regression Model

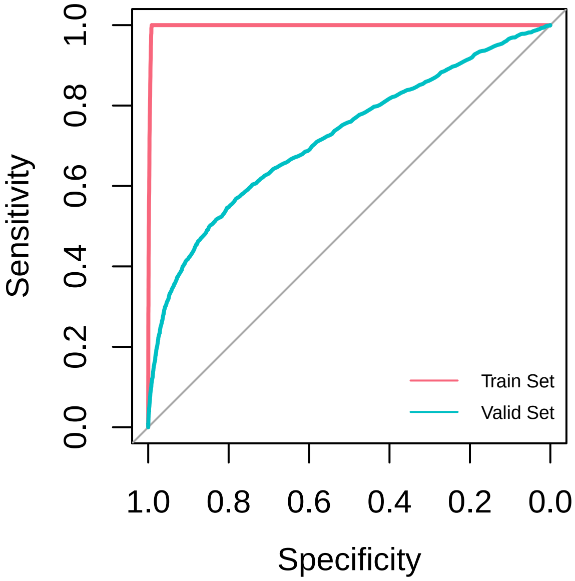

Figure 6: ROC Plot For Training and Validation Datasets of Random Forest Model

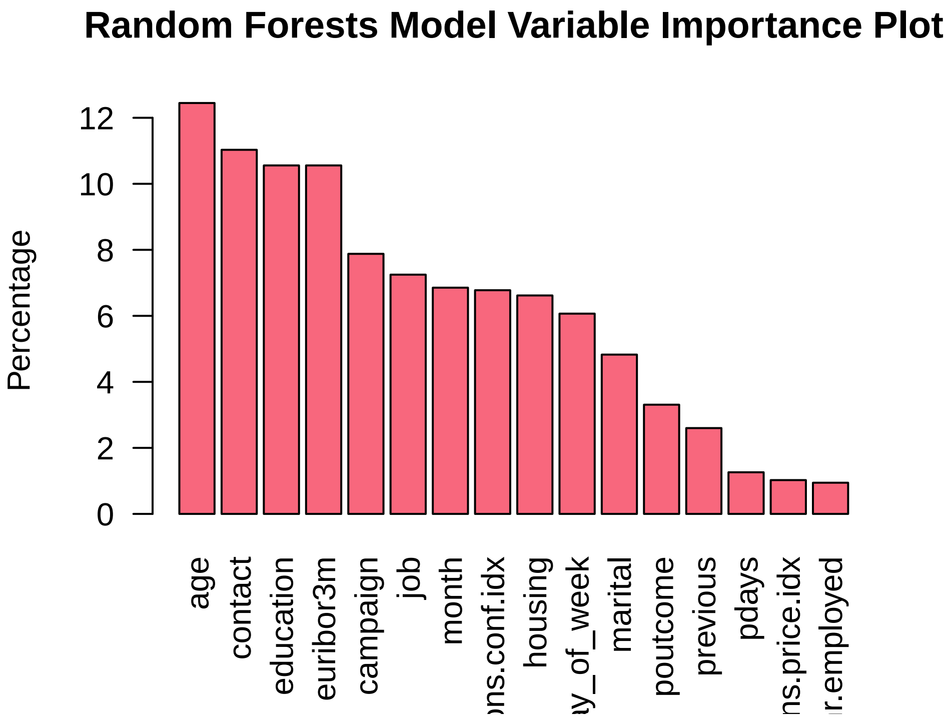

Figure 7: Plot of Random Forests Model Variable Importance